УДК 004.896

Македонский А. М., *Аксёнов К. А.*
УрФУ, г. Екатеринбург, Россия

ПОСТРОЕНИЕ ДЕРЕВЬЕВ ПРИНЯТИЯ РЕШЕНИЙ С ИСПОЛЬЗОВАНИЕМ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ

Аннотация

При работе с очень большими наборами данных построение дерева решений с использованием всех примеров может быть затруднительно. Даже в случаях, когда это возможно, это может быть не лучший способ использовать данные. В качестве альтернативы исходный набор данных может быть разделен на выборки, дерево может быть построено на каждом подмножестве, а затем отдельные деревья могут быть интеллектуальным образом объединены для получения окончательного оптимизированного дерева. В этой статье мы предлагаем случайным образом разделить исходные данные на выборки, произвести скрещивание и мутацию (с использованием генетического алгоритма) для ряда поколений выборок, чтобы получить более точные деревья.

Ключевые слова: деревья решений, генетический алгоритм, система поддержки принятия решений, алгоритм CART, алгоритм C4.5.

Makedonsky A. M., *Aksyonov K. A.*
UrFU, Ekaterinburg, Russia

GENETIC ALGORITHM-BASED APPROACH FOR BUILDING DECISION TREES

Abstract

In dealing with a very large data set, it might be complex to construct a decision tree using all of the examples. Even when it is possible, this might not be the best way to utilize the data. As an alternative, subsets of the original data set can be extracted, a tree can be constructed on each subset, and then individual trees can be combined in a smart way to produce an improved final tree. In this paper, we suggest dividing data set to randomly generated subsets and allowing them to crossover and mutate (using a genetic algorithm) for a number of generations in order to yield more accurate trees.

Keywords: Decision trees, Genetic algorithm, decision support system, CART algorithm, algorithm C4.5.

Введение

Деревья решений — один из методов автоматического анализа данных. Первые идеи создания деревьев решений восходят к работам Ховленда и Ханта конца 50-х годов XX века. Однако основополагающей работой, давшей импульс для развития этого направления, явилась книга Ханта, Мэрина и Стоуна «Experiments in Induction», увидевшая свет в 1966 г.

1. Построение деревьев

Пусть нам дано некоторое обучающее множество $S = s_1, s_2, \ldots$ уже классифицированных примеров, каждый из которых представляет собой вектор с p параметрами, причем один из них указывает на принадлежность примера к определенному классу.

Идею построения деревьев решений из множества S, впервые высказанную Хантом, приведем по Р. Куинлену [1].

Пусть через $\{C_1, C_2, \ldots C_k\}$ обозначены классы, тогда существуют три ситуации:

множество S содержит один или более примеров, относящихся к одному классу C_k. Тогда дерево решений для S — это лист, определяющий класс C_k;

множество S не содержит ни одного примера, т. е. пустое множество. Тогда это снова лист, и класс, ассоциированный с листом, выбирается из другого множества, отличного от S, скажем, из множества, ассоциированного с родителем;

множество S содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество S на некоторые подмножества. Для этого выбирается один из признаков, имеющий два и более отличных друг от друга значения $O_1, O_2, \ldots O_n$ и S разбивается на подмножества $S_1, S_2, \ldots S_n$, где каждое подмножество S_i содержит все примеры, имеющие значение O_i для выбранного признака. Это процедура будет рекурсивно продолжаться до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу.

Вышеописанная процедура лежит в основе многих современных алгоритмов построения деревьев решений, этот метод известен еще под названием «разделяй и властвуй» («divide and conquer»). Процесс, идущий «сверху вниз», индукция деревьев решений (TDIDT) [2], является примером поглощающего «жадного» алгоритма и на се-

годняшний день является наиболее распространенной стратегией построения деревьев решений, но это не единственная возможная стратегия. В интеллектуальном анализе данных деревья решений могут быть использованы в качестве математических и вычислительных методов, чтобы помочь описать, классифицировать и обобщить набор данных, которые могут быть записаны следующим образом: $(x, Y) = (x_1, x_2, \ldots x_k, Y)$. Зависимая переменная Y является целевой переменной, которую необходимо проанализировать, классифицировать и обобщить. Вектор x состоит из входных переменных x_1, x_2, x_3 и т. д., которые используются для выполнения этой задачи.

На сегодняшний день существует значительное число алгоритмов, реализующих деревья решений: CART, C4.5, NewId, Itrule, CHAID, CN2 и т. д. Но наибольшее распространение и популярность получили следующие два [3]:

CART (Classification and Regression Tree) — это алгоритм построения бинарного дерева решений — дихотомической классификационной модели. Каждый узел дерева при разбиении имеет только двух потомков. Как видно из названия алгоритма, решает задачи классификации и регрессии. Авторы — Брейман, Фридман, Стоун и Олшен;

C4.5 — алгоритм построения дерева решений, где количество потомков в узле не ограничено. Не умеет работать с непрерывным целевым полем, поэтому решает только задачи классификации. Автор: Дж. Р. Куинлен.

Для построения дерева на каждом внутреннем узле необходимо найти такое условие (проверку), которое бы разбивало множество, ассоциированное с этим узлом на подмножества. В качестве такой проверки должен быть выбран один из атрибутов. Общее правило для выбора атрибута можно сформулировать следующим образом: выбранный атрибут должен разбить множество так, чтобы получаемые в итоге подмножества состояли из объектов, принадлежащих к одному классу, или были максимально приближены к этому, т. е. количество объектов из других классов («примесей») в каждом из этих множеств было как можно меньше.

Алгоритм C4.5, усовершенствованная версия алгоритма ID3 (*Iterative Dichotomizer*), использует теоретико-информационный подход. Для выбора наиболее подходящего атрибута, предлагается следующий критерий:

$$Gain(X) = Info(S) - Info_x(S) \qquad (1)$$

где, Info (S) — энтропия множества S, а

$$Info_x(S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Info(S_i). \quad (2)$$

Множества S_1, S_2, ... S_n получены при разбиении исходного множества S по проверке X. Выбирается атрибут, дающий максимальное значение по критерию (1).

Впервые эта мера была предложена Р. Куинленом в разработанном им алгоритме ID3. Кроме вышеупомянутого алгоритма C4.5, есть еще целый класс алгоритмов, которые используют этот критерий выбора атрибута.

Алгоритм CART использует так называемый индекс Джини (Gini), который оценивает «расстояние» между распределениями классов:

$$I_G(f) = 1 - \sum_{i=1}^{m} f_i^2, \quad (3)$$

где f — текущий узел, а f_i — вероятность класса i в узле f.

На практике в результате работы этих алгоритмов часто получаются слишком детализированные деревья, которые при их дальнейшем применении дают много ошибок. Это связано с явлением переобучения. Для сокращения деревьев используется «отсечение ветвей» (pruning). Качество классификационной модели, построенной при помощи дерева решений, характеризуется двумя основными признаками: точностью распознавания и ошибкой. Точность распознавания рассчитывается как отношение объектов, правильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении. Ошибка рассчитывается как отношение объектов, неправильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении. Отсечение ветвей или замену некоторых ветвей поддеревом следует проводить там, где эта процедура не приводит к возрастанию ошибки.

2. Использование генетических алгоритмов

Для улучшения качества классификации, распознавания и прогнозирования предлагается использовать генетический алгоритм. В качестве популяций используются выборки из обучающего множе-

ства. Изначально все примеры имеют одинаковую вероятность попадания в выборку. На этапе скрещивания алгоритм повышает вероятность использования примеров из выборок, на основе которых получены наиболее точные деревья решений. Для оценки точности построенного дерева используется полное множество примеров. Деревья могут быть получены по одному или нескольким методам, при этом генетический алгоритм не привязан к алгоритмам построения дерева и пригоден для любых условий. Процедура мутации случайным образом исключает или добавляет примеры в выборку.

Работа алгоритма завершается после выполнения заданного количества этапов эволюции или при получении дерева с заданной точностью. На последнем этапе мы получаем некий лес деревьев решений с возможностью коллективной классификации и прогнозирования. Используя метод голосования, примеру приписывается тот класс, которому отдает предпочтение большинство деревьев из набора.

Аналогичный метод генерации деревьев решений проходит проверку в системе поддержки принятия решений для металлургических предприятий, создаваемой в рамках проекта 2012–218–03–167.

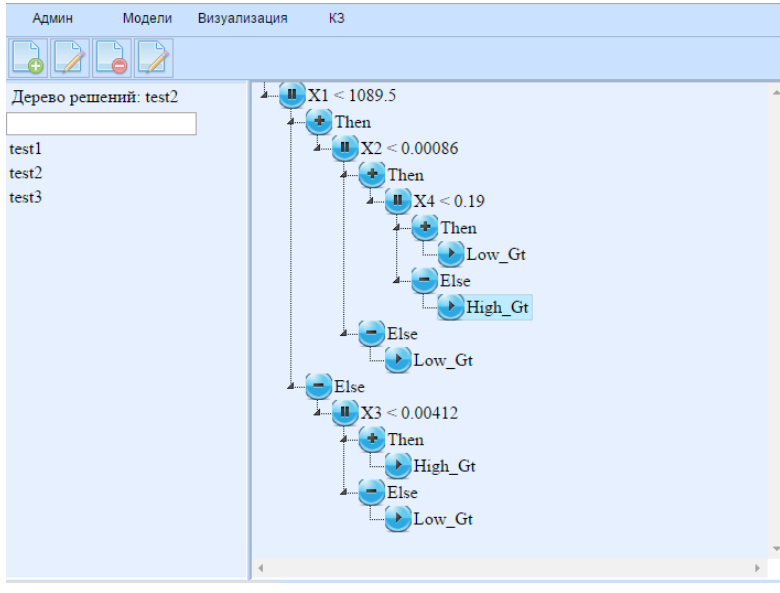

Рис. 1. Модуль генерации, редактирования и тестирования деревьев решений

Для генерации деревьев решений выбран алгоритм CART, т. к. он не подвержен влиянию аномальных данных, полученных в исключительных ситуациях, и обладает рядом других преимуществ. Большой объем данных, получаемый на предприятии в режиме реального времени, не позволяет строить наглядные и эффективные деревья без процедуры пост-обработки. Предлагаемый метод конкурирующих выборок позволяет получить более точные деревья решений, не требующие длительной процедуры отсечения ветвей.

3. Заключение

Деревья решений обладают рядом преимуществ, делающим их важным инструментом в интеллектуальном анализе данных:

быстрый процесс обучения;

генерация правил в областях, где эксперту трудно формализовать свои знания;

извлечение правил на естественном языке;

интуитивно понятная классификационная модель;

высокая точность прогноза, сопоставимая с другими методами;

построение непараметрических моделей.

При этом существует проблема построения деревьев решений. Несмотря на широкое применение, эвристические методы построения деревьев решений принципиально не способны находить наиболее полные и точные правила классификации данных, поскольку они основаны на использовании жадных алгоритмов. Предложенный способ оптимизации должен частично решать эту проблему без существенного усложнения алгоритмов построения.

Работа выполнена в рамках договора № 02.*G25.31.0055 (проект 2012–218–03–167)* при финансовой поддержке работ Министерством образования и науки Российской Федерации.

Список литературы

1. Quinlan J. R. C4.5: Programs for Machine learning // Morgan Kaufmann Publishers. 1993.

2. Quinlan, J. R. Induction of Decision Trees // Machine Learning. Kluwer Academic Publishers. 1986. № 1. P. 81–106.

3. Шахиди А. Деревья решений — общие принципы работы. URL: http://www.basegroup.ru/library/analysis/tree/description/ (дата обращения: 15.03.2015).